Accepted for publication in the Proceeding of *IS&T and SPIE Storage and Retrieval for Media Databases 2000*.
San Jose,California,USA. 23-28 January,2000.

Apply Semantic Template to Support Content-based Image Retrieval

Yueting Zhuang[※§] Xiaoming Liu[※§] Yunhe Pan[※§]

Institute of Artificial Intelligence, ZheJiang University[※]
Microsoft Visual Perception Laboratory of Zhejiang University[§]
HangZhou, 310027, P.R.China

Email:{ Yzhuang, Liuxm }@icad.zju.edu.cn , Panyh@sun.zju.edu.cn

ABSTRACT

Content-based multimedia information retrieval is the hot point of researchers in many domains. But traditional feature vector based retrieval method can not provide retrieval on the semantic level. Integrated with our image retrieval system, we propose a new approach to generate semantic template automatically in the process of relevance feedback, and construct a network of semantic template with the support of WordNet™ in the retrieval process, which helps the user to do retrieval on the semantic level. By our approach, in the keyword query of the user, relevant images will be returned to the user by the help of semantic template association even those images are not annotated by keyword. This paper introduces this approach in detail and presents an experiment result at the end of this paper.

Keywords: Multimedia, Information retrieval, Relevance feedback, Semantic template

1. INTRODUCTION

With the flourishing of multimedia application and the advancement of computer technique, content-based multimedia information retrieval has been the hot point of researchers in many research domains. Begin from the annotation of keyword by human labor to feature extraction automatically, great progress has been achieved in multimedia information retrieval. Many research prototypes and commercial systems have been proposed, such as Virage[3], VisualSeek[5], MARS[4]. For the example of image, content-based retrieval is on the basis of extracted feature vector from the image. So in the early, many researchers in the computer vision domain focus on how to find the "best" representation of image feature, which can include enough image information in the vector. But by the further research, people find that there are still not relevance images to the query in the list of result images although feature vectors match very well. Such as, system return an image includes a red apple while user submits a sunrise image as a example query. The main reason for this problem is:

- The method of image indexing based on the feature of low level can not represent the semantic of image effectively. Though user has so many approaches to extract the image feature, there is always a gap between the feature vector and the semantic. The system can not deduce semantic from the feature vector.
- The similarity between images is a concept associated with the subjective judgement of human. Different people use different criterions while in judgement. Even the same people use different criterion arming to different things. For example, one will usually pay more attention on the color when judging the similarity of two flowers, yet on the texture when judging two upholster wallpapers. At present, most systems use a uniform similarity measure, which can not embody the subjectivity of human.

On the other hand, most systems support the method of example query. But what can the user do when he does not have an example image at all? In a system used the keyword query, all the images needed to be annotated by text, which is implemented by the human labor on most situations.

All these questions presented in all content-based retrieval systems. Many researchers have been aware of it and do some work on it. Vailaya[10] show how a specific high-level classification problem (city images vs. landscapes) can be solved

from relatively simple low-level features geared for the particular classes. This means in the pattern recognition is usable in reducing the scope of retrieval. But the catalogues it can classificate are very finite. Picard[6] names some predefined feature model as a society model, then selects model which is most appropriate to the user's retrieval via relevance feedback. Rui[8] utilizes interactive retrieval manner to adjust the weight in the similarity measure automatically based on relevance feedback, such that the adjusted query is a better approximation to the user's information need. S.f.Chang[2] proposes a set of icons to represent the semantic visual template which link the low level feature to high level semantic. But their method of template generation depends more on the interaction of user and requires the user's in-depth understand of interior feature representation. So it is not applicable to the ordinary user.

Integrated with our image retrieval system, we propose a new approach to generate semantic template automatically in the process of relevance feedback, and construct a network of semantic template in the retrieval process, which helps the user to do retrieval on the semantic level. This paper is organized as follows. Section 2 introduces the architecture of our image retrieval system. The generation of semantic template is detailed in section 3. In section 4 we show how to use WordNet to associate semantic template. Section 5 shows the experiment result. Finally we give the conclusion and future directions.

2. THE ARCHITECTURE OF SYSTEM

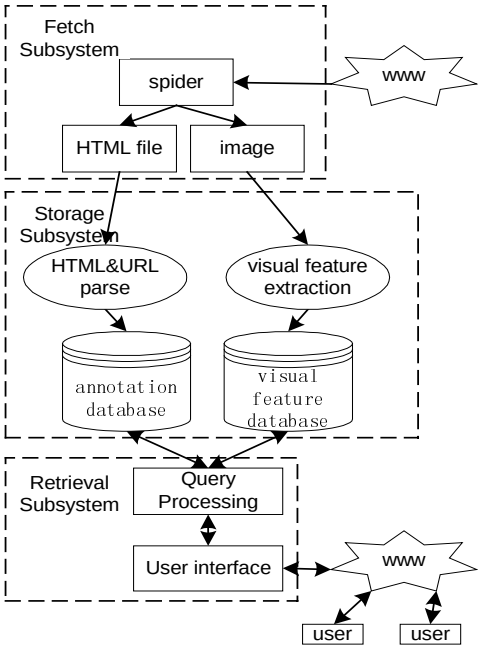

Fig.1. The architecture of our system

The architecture of our system is shown in figure 1. In the fetch subsystem, we designed a spider to download all the web pages contained images begin with a given URL. Because nearly all the web pages include images, we skim the image whose width and height are smaller than 120 pt. In addition, only the URL and representative icon of a image are saved in our local database.

In the storage subsystem, we implement the automatic annotation of image by two approaches. Firstly, the URL of many image files have a clear hierarchical structure, which contain some information about this image. So by the help of a predefined catalogue tree we parse the URL string and get which catalogue this image pertains to. Secondly, in the HTML file, there is usually a paragraph of descriptive text near the image location, which gives some general information about this image. As to it, we utilize traditional information retrieval technique to extract some useful information and regard them as the annotation of this image. As known from it, these two approaches can only annotate some image a certain extent. Our later work is also based on this premise.

We denote the visual feature of an image by:

$$F=\{f_i\}=\{f_{ij}\} \quad (1)$$

Where f_i represent a sort of feature representation, f_{ij} represent the j element of feature representation f_i. Presently the visual features we utilize are color(f_1) and texture(f_2). As to color feature, we convert every pixel from *RGB* space to *HSV* space, then calculate histogram using *H* and *S* weights in a two-dimension space, and finally regard 32 normalized values as features. We use the coarseness, contrast and direction defined by Tamura to represent texture feature, which is a vector consist of three values. Conclusively, the visual feature of every image is represented as a vector, which is composed of 35 values.

3. THE GENERATION OF SEMANTIC TEMPLATE

Semantic template is a map between semantic concept of high level and visual feature of low level. It is a deduction on the two directions. In current feature vector based image retrieval system, when user enter a keyword (concept), the system like to know the feature vector corresponding to this concept, thus the similar image can be found in the vector space even it is not annotated by the system. On the other hand, given a feature vector, the work of image understanding is implemented if the system can induce the semantic of this vector. So semantic template is significative for not only the retrieval of semantic level but also the recognition of image content.

3.1 The generation method integrated with relevance feedback

Comparing to the method used in S.f.chang[2], our approach integrates the generation process with the interactive relevance feedback. Anyone can use it conveniently even he has not any knowledge about the feature representation, and the user is imperceptible of the template generation. Relevance feedback is a powerful technique used in traditional text-based information retrieval system. It is the process of automatically adjusting an existing query using the information feedback by the user about the relevance of previously retrieved objects. In fact, it is also a process by which user embody the semantic better by the retrieval result. On one hand, the result set annotated as high relevance has the representative images of user's concept. On the other hand, this process makes the system to find which feature represent or element is the main criterion of user's similarity judgement. Given these two points, we think it is possible to integrate the generation of semantic template with relevance feedback (shown as figure 2). Firstly, when an annotated image is submitted to the system as a query example, the user also submits a concept represented by this image. The query processing return the list of result image based on the calculation of initial weight and visual feature vector. Then user annotates the relevance measure for every result image according to his comprehension of semantic concept. The query processing updates the weight based on the relevance feedback and calculates again. After several interactions between human and retrieval system, the most relevant images will be returned to the user. At last, the system calculates the centroid of those high ones and regards this centroid vector as the representation of the high semantic concept. Thus the semantic template is obtained. We define the semantic template as:

$$ST=\{C,F,W\} \tag{2}$$

Where C represent the concept of the user, $F=\{f_{ij}\}$ is the feature vector correspond to C, $W=\{W_i,W_{ij}\}$ is the weight of feature vector. To include W in the semantic template is based on the idea that people have different preference to the feature representation or element while regarding distinct concept, such as the flower and wallpaper mentioned before. So this triple represents the concept of semantic template completely, which can be used in retrieval directly. Now we introduce the relevance feedback technique used in our system.

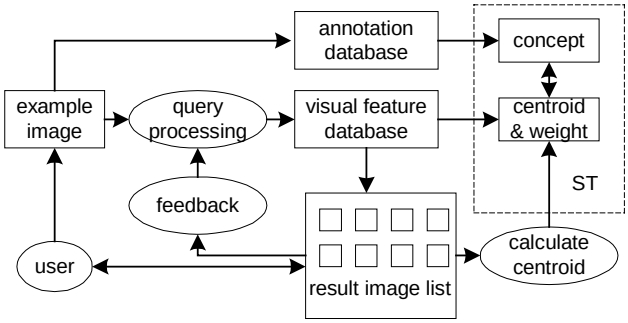

Fig. 2. The generation of ST integrated with relevance feedback

3.2 Relevance feedback based image retrieval technique

Given two images, I_1 and I_2, we calculate their similarity as below. Suppose the feature vector of these two images are F^1,F^2. We convert the final similarity value in the range [0,1]. 1 means they have the most similarity, and 0 means they do not similar entirely. Because the attributions of two vectors composed the visual feature vector are different, we should distinguish them while calculating the vector distance. As to the color vector, we can use Euclidean distance as below directly for it has been normalized in the range [0,1] and defined over the same physical domain.

$$Similarity_{fi} = \sum_{j=0}^{31} W_{1j} \bullet Min(f_{1j}^1, f_{1j}^2) \tag{3}$$

As to the vector-based feature representation, such as texture, every value does not in the same range. So we should convert all these values into the same range before the utilization of distance measure. By the example of f_{21} in texture feature, let us see how to use Gaussian normalization method. Suppose there are k images, denoted as $f_{21}^{\ i}$ $(i=1\sim k)$, in image database. We calculate the means m and difference σ of $f_{21}^{\ i}$ firstly, and then normalized k values into range [-1,1] by

$$f_{21}^{\ i}{}' = \frac{f_{21}^{\ i} - m}{3\sigma} \qquad i=0,..,k \tag{4}$$

As known from Gaussian distribution, the probability of $f_{21}^{\ i}$ in the range [-1,1] is approximately 99%. So we should map the out-of-range values to either –1 or 1. The normalization of *coarseness* and *direction* is same as before. Then we can accumulate the distance of three values as the texture vector distance.

Now the respective distances of two vectors have been achieved. But only the color distance is in the range [0,1]. So we have to normalize the texture distance by Guassian normalization also. But after using formula like (4) to get $similarity_{f2}$, the below liner transformation is needed for converting it to the range [0,1].

$$Similarity_{f2} = 1 - \frac{Similarity_{f2} + 1}{2} \tag{5}$$

At last, we can get the similarity between two images according to two weights W_1,W_2. as below.

$$Similarity(I_1,I_2) = W_1 \bullet Similarity_{f1} + W_2 \bullet Similarity_{f2} \tag{6}$$

Then for any image submitted by the user, our system can calculate its similarity with every image in database, and return the most similar N images. The W_i and W_{ij}, reflect the user's different emphasis of a feature vector. The support of different weights enables the user to specify his concept more precisely.

Let R be the set of the most similar N images according to the overall similarity value similarity.

$$R=[R_1,R_2,\ldots,R_k,\ \ldots,R_N] \tag{7}$$

For every R_t in set R, the system use scroll bar to let user feedback his relevance judgement, $Score_k$. The most right end of the scroll bar represent the highest relevant($score=1$); the most left end represent the highest non-relevant($score=-1$); the middle point represent that the user has no opinion about the relevance($score=0$); other location can be set a liner value. Thus, $score_k$ is defined as a float point in a one-dimensional scope of (-1,1).

For each f_i, let R^i be the set containing the most similar N images according to the similarity value, $Similarity_{fi}$:

$$R^i=[R_1^{\ i},R_2^{\ i},\ldots,R_k^{\ i},\ldots,R_N^{\ i}] \tag{8}$$

Then the algorithm shown in figure 3 is used to update W_i. Comparing with the algorithm in Rui[8], we consider not only the overlap of relevant images between R and R_i, but also the similarity between the order of element in R_i and the order of user's relevant value in R. So it can embody the user's preference better.

When updating W_{ij}, suppose there are m result whose relevant value are bigger than 0. We stack these m feature vectors to form a $m \times k$ matrix, and calculate the standard deviation, σ_k, of every column. If the values in a column are very similar, it means that column has a more contribution to the overall similarity. So the smaller the variance, the larger the weight and vice versa.

$$W_{ij}=1/\ \sigma_k \tag{9}$$

Then use the follow formula to update W_{ij}:

$$W_{ij} = \frac{1}{\sum W_{ij}} \tag{10}$$

Now by inputing the updated W_i and W_{ij} to the above similarity algorithm again, the system will return more relevant images to the user.

```
UpdateWi(R, R^i , Score)
Input:R：  The result image list by the similarity of all
the feature vector
R^i : The result image list by the similarity of the feature
vector f_i
Score: The feedback of user to the R set
Output: W_i: The updated weight
(1)   SumOfWi=0;
(2)   LastScore=0;
(3)   For ( i=1 ; i<=2 ; i++)
(4)   {  W_i=0;
(5)      For(k=1 ; k<=N ; k++)
(6)      {     j=0;
(7)            while( R_k^i<>R_j && j<=N )  j++;
(8)            If (R_k^i==R_j )
(9)            {     W_i=W_i+Score_j ;
(10)                 If( Score_j<LastScore )
(11)                        W_i=W_i+1/(N-1);
(12)                 LastScore=Score_j ;
(13)           }
(14)      }
(15)      If ( W_i<0 )  W_i=0;
(16)      SumOfWi= SumOfWi+W_i;
(17)  }
(18)  For ( i=1;i<=2;i++ )       W_i=W_i/ SumOfWi
```

Fig. 3. The algorithm of updating W_I

3.3 The generation of semantic template

After several feedbacks, the system adjusts the weight and results in more relevant images returned to the user. In fact, these image set is a good representation of the concept submitted by the user, and W_i and W_{ij} also embody the user's preference while consider the concept. So, based on the set, we generate semantic template like this:

1 Combine current values of W_i and W_{ij} to a vector:

$$W=\{W_{11},W_{12},\ldots,W_{132},W_{21},W_{22},W_{23}\} \tag{11}$$

2 In set R, regard all the images, R_k, whose relevant value is bigger than 0 to form a relevant set Ω, and calculate:

$$SumScore = \sum_{k\in\Omega} Score_k \tag{12}$$

3 Calculate:

$$f_{ij} = \{\frac{Score_k}{SumScore} \bullet f_{ij}^k \mid \forall k \in \Omega\} \tag{13}$$

4 Combine the concept of the user, C, with F and W: $ST=\{C,F,W\}$, where $F=\{f_{ij}\}$

This process is equal to regard every image as a point in the space of feature vector, whose mass is the relevant value feedbacked by the user. Step 3 calculate the centroid of all feature vectors whose corresponding image has a relevant value

bigger than 0. The centroid is the feature vector *F* in the semantic template. At present, we can generate many semantic templates for different concepts, but all these templates are isolated. Suppose there is a semantic template whose concept is "sunrise", it can not take effect when the user types in "natural scene " as keyword. So we need to link all the semantic templates on the semantic level. Thus they can be contributing in retrieval process. WordNet just implements this task.

4. USING WORDNET TO ASSOCIATE SEMANTIC TEMPLATE

WordNet is an electronic lexical system developed by George Miller and his colleagues at Princeton University. It is a lexical inheritance system. For example, the lexical tree can be reconstructed by following trails of terms: oak→tree→ plant(shown in figure 4). The noun portion of WordNet is designed around the concept of synset, which is a set of closely related synonyms representing a word sense (i.e. meaning). Every word that is in the WordNet has one or more senses and for each sense it has a distinct set of synonyms, and a distinct set of words related through other relationships such as IS_A relation, MEMBER_OF relation and PART_OF relation. In our system, we have used the noun, adverb and verb portions of WordNet. WordNet uses three ways to construe the semantic associations.

- The three relationships in noun portion: For example, ''man" and ''gentleman" is linked by IS_A. If the transcribed text in the database is ''man", while the term in user's query is ''gentleman", WordNet provides such kind of semantics association.
- Causal relation between actions: WordNet also provides ''Cause-to" relation between actions(verb) which can be used in the explanation of actions. For example: ''kill" causes ''die".
- Synonyms between adjectives: For example, looking for similar words to ''fantastic" is necessary. WordNet provides synonyms between adjectives. For example, similar words to ''beautiful" are ''attractive", ''charming", ''fine-looking", ''pretty" and so on.

What we do is to define a distance for the semantic association and a distance threshold *h* in wordnet. When the user type a keyword, the system find all the semantic link whose distance is smaller than *h*. Thus a ordered terms list is obtained according to the distance measure. At last, for every term in the list, the system finds its corresponding semantic template, and uses the *F* and *W* to query similar images. In fact, existing semantic template can make up a semantic network via wordnet, which confirm that system can provide semantic association on the condition of limited semantic template. That is, as soon as the user types in the keyword which have semantic link with existing semantic template, that ST can take function well.(shown in figure 5)

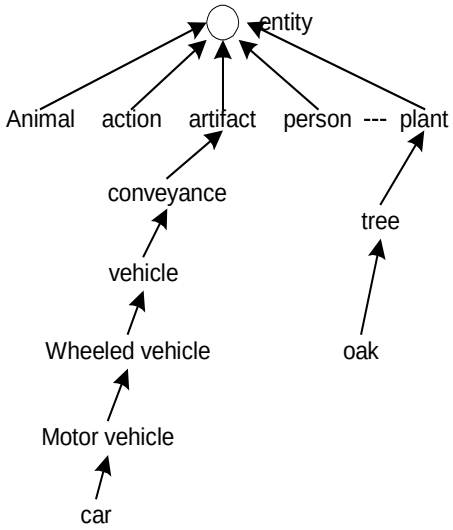

Fig. 4. The lexical tree of WordNet

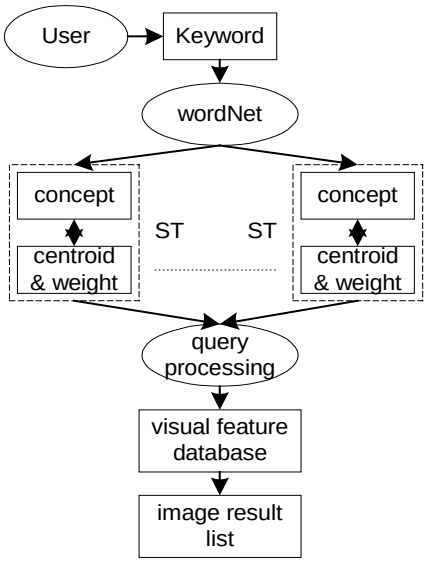

Fig. 5. Image retrieval supported by WordNet and ST

5. EXPERIMENT RESULTS

Recently, We have implemented an image retrieval system, *ImageSeek*, on personal computer. Based on the SQL Server database, the storage subsystem is implemented with visual C++, and the retrieval subsystem is done with Active Server Pages. Our image database consisted of 500 various images whose category include human, animal, car, natural scene, etc.

Let us see the generation of a semantic template. In the database there are nine images about “lawn”, within which only two images are annotated automatically. As shown in figure 6, we submit the left_top image, which is annotated as “lawn” in database, as query example. The system returns ten results according to the similarity order from left to right and top to bottom. Then the user annotates the relevant value for every result by the scroll bar. Based on it, the system updates the weights and returns more relevant results again(shown in figure 7). Through several interactions, the semantic template about “lawn” is generated. At last, we type in the word “meadow” as query keyword. With the support of WordNet and that semantic template, the system returns all the nine images about “lawn” in the first itineration.

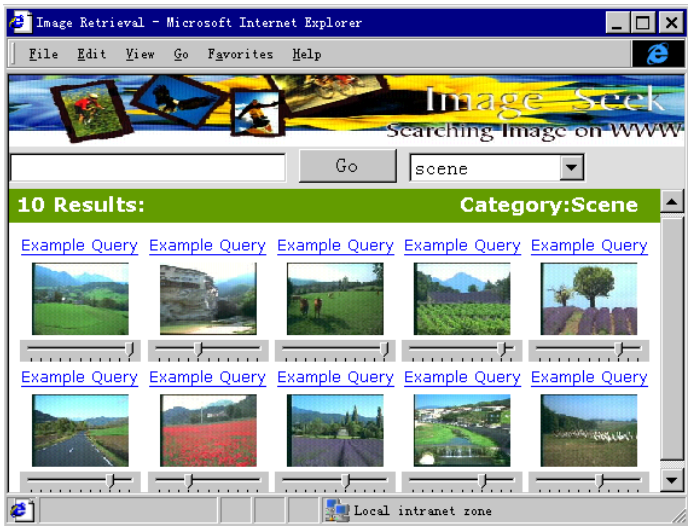

Fig. 6. The retrieval result of a submitted image example.

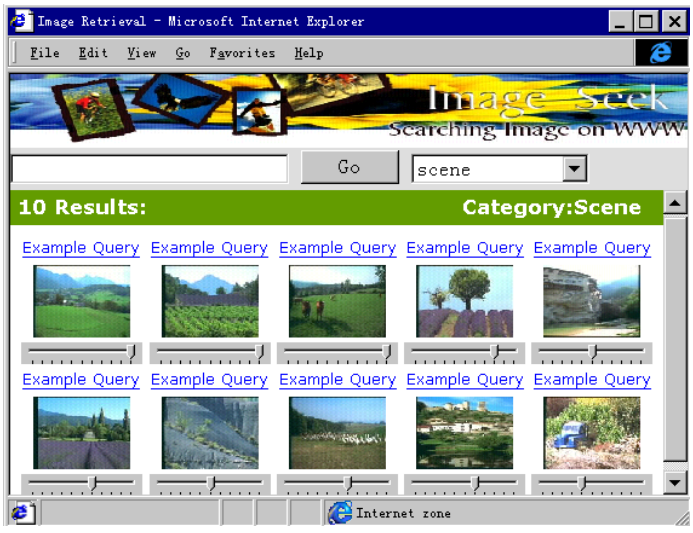

Fig. 7. The retrieval result after relevance feedback.

6. CONCLUSIONS

Content-based multimedia information retrieval is a hot point of researchers in many domains. But traditional feature vector based retrieval method can not provide retrieval on the semantic level. Integrated with our image retrieval system, we propose a new approach to generate *ST* automatically in the process of relevance feedback, and construct a network of *ST* with the support of WordNet in the retrieval process. By our method, many *ST*s are generated in the process of user's query by example image. In the keyword query of the user, relevant images will be returned to the user by the help of *ST* association even those images are not annotated by keyword. As known from it, this technique is general for any media retrieval although we utilize it in the image retrieval. In addition, any improvement in the technique of relevance feedback will embody in the generation of template, so this technique is open. Future work will concentrate on customizing personal template and merging multi-user template.

ACKNOWLEDGMENTS

Our work was sponsored by the National Natural Science Foundation of China. We would like to thank Yi Wu, Yi Mao for fruitful discussions.

REFERENCES

1. Y. Alp Aslandogan. et.al. Using Semantic contents and WordNet™ in image retrieval. Proceeding of ACM SIGIR 97, pp 286-295, Philadelphia PA,USA
2. Shih-Fu Chang, William Chen, Hari Sundaram. Semantic Visual Templates: Linking Visual Features to Semantics, ICIP'98, Workshop on Content Based Video Search and Retrieval, Chicago IL,Oct 10-12 1998
3. A. Hamrapur,et.al. Virage video engine. SPIE Proceedings on storage and retrieval for image and video databases V,pp 188-197, San Jose,Feb,1997
4. Sharad Mehrotra, Kaushik Chakrabarti, Mike Ortega, Yong Rui, and Thomas S. Huang, Multimedia Analysis and Retrieval System , Proc.of The 3rd Int. Workshop on Information Retrieval Systems , Como, Italy, September 25-27, 1997, pp39-45.
5. J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. ACM Multimedia'96, November, 1996.
6. R. W. Picard. A society of models for video and image libraries. MIT Media Lab. TR#360, Apr.1996
7. Yong Rui, Thomas S. Huang, and Shih-Fu Chang, Image Retrieval: Current Techniques, Promising Directions and Open Issues, Journal of Visual Communication and Image Representation , 10, 1-23,1999.
8. Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra, Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval , IEEE Tran on Circuits and Systems for Video Technology , Special Issue on Segmentation, Description, and Retrieval of Video Content, pp644-655, Vol 8, No. 5, Sept, 1998
9. Yueting Zhuang, Yong Rui, Thomas S. Huang, "Applying Semantic Association to Support Content-based Video Retrieval", Int.Workshop on Very Low Bitrate Video Coding (VLBV98),Oct.1998, USA.
10. A. Vailaya, A. K. Jain and H. J. Zhang, "On Image Classification: City Images vs. Landscapes" , Pattern Recognition, vol. 31, pp 1921-1936, December, 1998